DOCUMENT RESUME

ED 227 135 TM 830 129

AUTHOR Huberty, Carl J.
TITLE Discriminant Analysis.
PUB DATE Mar 82
NOTE 40p.; Paper presented at the Annual Meeting of the American Educational Research Assocation (66th, New York, NY, March 19-23, 1982).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Behavioral Science Research; *Classification; *Cluster Grouping; Computer Programs; *Discriminant Analysis; *Multivariate Analysis; Predictive Measurement; Regression (Statistics)

ABSTRACT

The issues in the interpretation of discriminant analyses presented are restricted to the typical uses of discriminant analysis by behavioral science researchers. Because behavioral researchers use computer programs packages, the issues discussed deal with information obtainable from the package disciminant analysis programs. The following issues are reviewed: (1) distinguishing between a linear discriminant function and a linear classification function; (2) misusing a stepwise discriminant analysis program; (3) ordering variables and selecting variable subsets; (4) choosing a classification rule; (5) estimating true classification hit rates; (6) assessing classification accurately; and (7) examining and using classification results. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED227135

Discriminant Analysis

Carl J Huberty

University of Georgia

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. J Huberty

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented in the symposium, "Issues in the Interpretation of Multivariate Analysis," at the Annual Meeting of the American Educational Research Association, New York, March 1982.

TM 830 129

Abstract

The two problems for which a discriminant analysis is used--discrimination and classification--are reviewed. Issues related to the interpretation of a discriminant analysis discussed are those pertaining to: (1) distinguishing between a linear discriminant function and a linear classification function; (2) misusing stepwise discriminant analysis programs; (3) ordering variables and selecting variable subsets; (4) choosing a classification rule; (5) estimating true classification hit rates; (6) assessing classification accuracy; and (7) examining and using classification results.

Discriminant Analysis

Overview 1

LDFs and LCFs 3

Stepwise Discriminant Analysis 7

- Statistics 8
- Uses 10

Ordering and Selecting Variables 11

- Ordering Indices 11
- Selection Methods 14
- Classification Analyses 16

Classification Hit Rates 17

- Rule Choice 17
- Internal and External Analyses 18
- Assessment of Observed Hit Rates 20
  - Proportional chance criterion 21
  - Maximum chance criterion 24
  - Separate-group hit rates 24
  - Improvement over chance 26
- Threshold Probabilities 28

Classification and Discrimination 29

Summary 30

Overview

There are two aspects of, i.e., ways of using, discriminant analysis. One aspect pertains to the study of the effect of the treatment (or experimental or grouping) variable and, if more than one, the joint effects of such variables. A preliminary analysis often conducted in such a case is a one- or higher-way multivariate analysis of variance (MANOVA), along with finding the associated P-values and variance-explained values. Specific interpretations and explanations of resulting effects may be accomplished by using a set of techniques collectively known as descriptive discriminant analysis (DDA). DDA techniques may be used to answer three questions: (1) Is there a subset of the original set of response variables that yields nearly the same effects as does the original set?; (2) What is a reasonable ordering of the retained set of variables in terms of their relative contribution to the resulting effects?; and (3) What is a reasonable substantive interpretation of the structure underlying the resulting effects?

There are requisite data conditions for conducting the MANOVA and the DDA. These are: (i) independence of observation vectors; (ii) multivariate normality of observation vectors in populations corresponding to "cells" in the design; and (iii) equal population covariance matrices. The two basic statistics that are appropriate under these conditions and useful in answering the three aforementioned questions are Wilks' lambda and the linear discriminant function.

The second aspect of discriminant analysis pertains to classifying units, on each of which are available multiple response measures, into

well-defined groups or populations. The setting here is that of a one-way layout. The collection of techniques used in classification problems is known as predictive discriminant analysis (PDA). PDA techniques refined to this date may be used to answer three questions: (1) What is the rule used to classify units?; (2) What are good estimates of separate-group and total-group percents of correct classifications (i.e., hit rates)?; and (3) Are the hit rate estimates better than those expected by "chance"? Two additional questions may be suggested: (4) Is there a subset of the original set of response variables that yields a hit rate as good as, or better than, that yielded by the original set?; and (5) What is a reasonable ordering of the retained set of variables in terms of their relative contribution to classification accuracy?

The only data condition truly necessary for conducting a PDA is independence of observation vectors. Most of the classification rules used in behavioral research, however, further assume that the multivariate normality condition is met. Since rules are available for use when covariance matrices are not equal, this condition is not a concern in PDA.

A basic PDA statistic is the linear classification function when covariance matrices are assumed equal, and the quadratic classification function when covariance matrices are not assumed equal. [In either case, an equivalent statistic, with respect to classification results, is an estimated posterior probability of group membership.]

It is true that a large majority of behavioral researchers use normal-based discriminant analysis techniques, in DDA and in PDA. This is undoubtedly because of computer programs available through the popular packages: SPSS, BMD, and SAS. Even though procedures for handling non-normal variables present new interpretation issues, there still remain a

number of issues pertaining to the interpretation of normal-based discriminant analyses. These are the issues to be discussed in this paper--some are of a general nature, some pertain to specific computer program output.

LDFs and LCFs

Consider a one-way layout with k groups and p response variables. Let E be the p x p error (or, within-groups) sums-of-squares and cross-products (SSCP) matrix, and H be the p x p hypothesis (or, between-groups) SSCP matrix. Then the set of $r = \min(p, k - 1)$ linear discriminant functions (LDFs) is found by finding the r v-vectors (i.e., eigenvectors) from

$$(E^{-1} H - \lambda I)\, \underline{v} = \underline{0}, \qquad [1]$$

where the λ-values are the r eigenvalues of $E^{-1} H$, and I is the p x p identity matrix. The jth LDF may be expressed as

$$Y_j = v_{j1} X_1 + \ldots + v_{jp} X_p.$$

[Except in the case of k = 2 and one LDF, there is no explicit expression for the p x 1 vector, $\underline{v}_j$.] It may be noted that such an eigenanalysis can be carried out without assuming or imposing any variable vector distributional form. That is, we need not assume multivariate normality to find LDFs, they are mere linear combinations of the p X variables. It is reasonable, however, in working with equation [1] that the population covariance matrices be assumed common, since S ($= \nu_e^{-1} E$, where ν_e is the error degrees-of-freedom value) is implicitly used as the common covariance matrix estimator.

In a higher-way layout, there is a set of LDFs for each effect; in an A x B layout, for example, there are three sets of LDFs, one for the A effect, one for the B effect, and one for the AB effect. The number of LDFs in each set is $r = \min(p, \nu_h)$, where ν_h is the number of hypothesis degrees of freedom. Just as in a one-way layout, each set of LDFs is determined using equation [1].

The elements of the v-vectors in [1], which are unique only up to a constant of proportionality, are outputted by SPSS's DISCRIMINANT program under the heading, UNSTANDARDIZED CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS, by BMD's P7M program under the heading, COEFFICIENTS FOR CANONICAL VARIABLES, and by SAS's GLM program under the heading, CHARACTERISTIC VECTOR.

A rationale is now provided that leads to the formulation of a linear classification function. The basic idea in a predictive discriminant analysis is to classify a unit with the p x 1 observation vector $\underline{X}$ into group G_g ($g=1,\ldots,k$) if the (posterior) probability of belonging to G_g, given its vector $\underline{X}$, is larger than the probability for any other group. In this way we are classifying the unit into that group to which it is closest, in a probabilistic sense. These posterior probabilities are functions of prior probabilities of group membership, $P(G_g)$, and of probability densities (or, likelihoods) for each population, $f(\underline{X}|G_g)$:

$$P(G_g|\underline{X}) = \frac{P(G_g) \cdot f(\underline{X}|G_g)}{\sum_{g'=1}^{k} P(G_{g'}) \cdot f(\underline{X}|G_{g'})} . \qquad [2]$$

It is obvious from [2] that, for group assignment decisions, one could focus on the numerator alone instead of on the whole ratio, i.e., focus on $P(G_g) \cdot f(\underline{X}|G_g)$. To find these values, one for each group for each unit,

it is necessary to specify some distributional form for the data; it is simpler to consider the same form for all populations. If f is assumed to be a multivariate normal density function, then maximizing $P(G_g) \cdot f(\underline{X}|G_g)$ is equivalent to maximizing

$$\ln P(G_g) - \tfrac{1}{2} (\underline{X} - \underline{\mu}_g)' \Sigma_g^{-1} (\underline{X} - \underline{\mu}_g) - \tfrac{1}{2} \ln|\Sigma_g|, \qquad [3]$$

where $\underline{\mu}_g$ is the mean vector for population g, and Σ_g is the covariance matrix for population g. Assuming equal covariance matrices, an estimator for [3] is

$$\ln q_g - \tfrac{1}{2} (\underline{X} - \underline{\bar{X}}_g)' S^{-1} (\underline{X} - \underline{\bar{X}}_g) - \tfrac{1}{2} \ln|S|, \qquad [4]$$

where $q_g = \hat{P}(G_g)$, S is an estimator for the common covariance matrix, and $\underline{\bar{X}}_g$ is an estimator for $\underline{\mu}_g$. Algebraic manipulation leads to a statistic that is classification-equivalent to [4], and thus to an estimator for [2], called a linear classification function (LCF):

$$L_g = (\underline{\bar{X}}_g' S^{-1}) \underline{X} - \tfrac{1}{2} \underline{\bar{X}}_g' S^{-1} \underline{\bar{X}}_g + \ln q_g . \qquad [5]$$

An LCF may be written as a linear combination,

$$L_g = b_{g1} X_1 + b_{g2} X_2 + \ldots + b_{gp} X_p + C,$$

where the 1 x p vector of weights is $\underline{b}' = \underline{\bar{X}}_g' S^{-1}$, and the constant is $C = -\tfrac{1}{2} \underline{\bar{X}}_g' S^{-1} \underline{\bar{X}}_g + \ln q_g$. The LCF in [5] is used in classification by assigning the unit to that group which yields the largest L_g value.

It should be noted that the only similarity between an LCF and an LDF is that they both are linear combinations of the response variables. For a one-way layout with k groups there are $r = \min(p, k - 1)$ LDFs and k LCFs, they are derived differently, and they serve different purposes.

The elements in the b-vectors and the C-constants of the k LCFs are outputted by SPSS's DISCRIMINANT program under the heading, CLASSIFICATION FUNCTION COEFFICIENTS (with an inappropriate subheading), and by BMD's P7M program under the heading, CLASSIFICATION FUNCTIONS; SAS's DISCRIM program does not output the LCFs.

Some behavioral researchers have expressed apparent confusion in their writings when they discuss the use of LDFs in classification. Classification decisions can be based directly on LDF scores only in a two-group situation. [This is true because the single LDF found when k = 2 has weights that are proportional to weights of $L_1 - L_2$.] It is true that in the SPSS Manual (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975, pp. 435, 455), use of LDFs in k > 2 classification problems is implied. What they are doing, however, is simply using LDF scores as input to LCFs. The final classification decision still depends on values of LCFs (or, equivalently, on estimated posterior probability values).

Now, if the k group covariance matrices are equal and if all r LDFs are used, classification results using LDF score input will be identical to those using raw score input. When the assumption of equal population covariance matrices is tenable, results of classification in r-dimensional spaces may be expected to be very close to those of classification in p-dimensional spaces. The reason some computer programs use LDF scores, rather than raw scores, as input is that it is usually computationally more efficient to classify in r-dimensional spaces than in p-dimensional spaces. Using raw scores as input, all vectors in [5] have p elements and S is p x p; using LDF scores as input, the vectors have r elements and S is r x r. Since, in most situations, r is much smaller than p, inverting an r x r matrix is more efficient than inverting a p x p matrix. [It should be kept in mind that the SPSS program determines LDFs for other purposes as well.]

One last comment about classifying in the r dimensional "discriminant spaces." Not assuming equal population covariance matrices the estimator for [3] is called a quadratic classification function (QCF):

$$Q_g = \ln q_g - \tfrac{1}{2} (\underline{X} - \bar{\underline{X}}_g)' S_g^{-1} (\underline{X} - \bar{\underline{X}}_g) - \tfrac{1}{2} \ln|S_g|. \qquad [6]$$

What the SPSS DISCRIMINANT program does (when Option 14 is specified) is, first transform the p response variable scores to r LDF scores; then it uses the r LDF scores as input for [6]. It should be noted that, even if all r LDFs are considered, the "quadratic" classification results from this SPSS analysis may be quite different than if raw score input is used in [6].

As noted above, in the two-group case, a classification rule may be based on $L_1 - L_2$ (or, equivalently, on the single resultant LDF). With this in mind, it might be reasonable to base a k-group classification rule on $L_g - L_{g'}$, $g \neq g' = 1, \ldots, k$. This is, in fact, what was proposed in the classical multivariate analysis text by Anderson (1958, pp. 147-150), and hence presented by Morrison (1976, pp. 239-242) and by Cooley and Lohnes (1971, pp. 268-269). [In all three references, the prior probabilities, $P(G_g)$, are ignored.] It may be shown, however, that the rule based on the $L_g - L_{g'}$ differences, ignoring priors, is equivalent to the rule based on the prior-less L_g values themselves (see Morrison, 1976, p. 241; Gondek, 1981).

Stepwise Discriminant Analysis

Without a doubt, the most popular way of "doing a discriminant analysis" is to run a stepwise discriminant analysis using one of the computer program packages. Two statistics are the focus of attention for most uses of stepwise analysis programs. These will be presented in some detail in this section, along with a recently proposed statistic. Following a discussion of the three statistics, uses (really, misuses) of stepwise discriminant analysis programs are presented.

Statistics

Values of two statistics are typically considered in a stepwise discriminant analysis: F-to-enter and F-to-remove. Suppose there are q (<p) variables already in the analysis. The F-to-enter value for X_i not in the analysis, F_i, is used to test the null hypothesis of the equality of the k means on X_i when the q variables, already in, are partialled out. This F_i is a transformation of a "partial Λ," $\Lambda(X_i|\underline{X})$, which reflects the separation[1] yielded by X_i in addition to that attributable to the q variables already in:

$$F_i = \frac{df_e - q}{k - 1} \cdot \frac{1 - \Lambda(X_i|\underline{X})}{\Lambda(X_i|\underline{X})}, \qquad [7]$$

where $\underline{X}$ is the vector of q measures already in. [Note that this is essentially an analysis of covariance.] A "large" F_i value, reference being made to an F distribution with df = $(k-1)/(df_e - q)$, indicates that X_i contributes something to group separation in addition to the q variables already in; i.e., if X_i is entered, then group separation would be significantly increased. [It should be noted that the WILKS' LAMBDA next to each F-to-enter on an SPSS DISCRIMINANT print-out is not a partial Λ, but the lambda that would result if the corresponding variable would be entered into the analysis; i.e., it is $\Lambda(\underline{X}, X_i)$.]

Again, suppose q (< p) variables are in the analysis. The F-to-remove value for X_i', F_i', is used to test the null hypothesis of the equality of the k means on X_i, when the other q - 1 variables are partialled out. An F_i, is a transformation of a partial Λ which reflects the separation yielded by X_i, conditioned on the removal of the other q - 1 variables:

$$F_{i'} = \frac{df_e - q + 1}{k - 1} \cdot \frac{1 - \Lambda(X_{i'}|\underline{X}^*)}{\Lambda(X_{i'}|X^*)} , \qquad [8]$$

where $\underline{X}^*$ is the vector of q - 1 measures. A "large" $F_{i'}$ value, reference being made to an F distribution with df = $(k - 1)/(df_e - q + 1)$, indicates that $X_{i'}$ does contribute something to separation in addition to the other q - 1 variables; i.e., if $X_{i'}$ is removed, then group separation would be significantly decreased. [Again, it should be noted that the WILKS' LAMBDA next to each F-to-remove on an SPSS DISCRIMINANT print-out is not a partial lambda, but the lambda that would result if the corresponding variable were deleted; i.e., it is $\Lambda(\underline{X}^*)$.]

These two statistics are discussed by Afifi and Azen (1979, pp. 310-318), Klecka (1980, pp. 57-58), and Jennrich (1977); another statistic, "tolerance," is also discussed by these writers.

There is a recent SAS update that includes a stepwise discriminant analysis program, STEPDISC. In addition to the F-to-enter and F-to-remove statistics, STEPDISC outputs a third statistic called a "squared partial correlation." Again, suppose there are q(< p) variables already in the analysis. The partial-R^2 value for X_i not in the analysis is an index of relationship between the grouping variable and X_i when the q variables are partialled out of X_i. As such, it is an η^2-like index; using the notation for the F-to-enter,

$$R_i^2 = 1 - \Lambda(X_i|\underline{X}) . \qquad [9]$$

A "large" R_i^2 value, equivalent to a "large" F_i value (see [7]), indicates that the grouping variable accounts for a significant proportion of the residual variance of X_i. The partial-R^2 value for $X_{i'}$, already in the analysis, $R_{i'}^2$, , reflects the proportion of variance of $X_{i'}$, conditioned on

the other q - 1 variables that may be attributed to the grouping variable:

$$R^2_{i'} = 1 - \Lambda(X_i, |\underline{X}^*) \quad . \qquad [10]$$

Uses

A reader of most, if not all, reports of studies that employ a stepwise discriminant analysis might conclude that a stepwise analysis may be used for two purposes: variable ordering and variable selecting. Typically, the applied researcher is interested in determining which variable (in the company of the remaining p - 1 variables) contributes most to group separation, which variable contributes next most, down to which variable contributes least. And, to do this, he/she often uses the order in which variables enter the analysis, based on values of the F-to-enter in [7], or, equivalently, the partial-R^2 in [9].

A stepwise analysis is also, typically, used to select a subset of q (< p) variables to retain for subsequent interpretation. To do this, the applied researcher often determines q by the magnitude of the F-to-enter in [7]; if no F_i value reaches some cut-off value (e.g., 2.0), then the subset of variables already in the analysis at the previous step is the retained subset.

It is unfortunate, I think, that applied researchers have been mislead into thinking not only that a stepwise analysis can yield such variable ordering and selecting information, but that a stepwise analysis is a "good," some even say the "best" (!), way to order and select variables. An attempt will be made in the next major section of the current paper to convince the applied researcher that stepwise discriminant analysis results should not be used for either purpose, ordering or selecting variables.

It is important to note that the three statistics and the two uses of stepwise analyses discussed above should be considered only in the context of a group separation, or discrimination, problem, if at all. These statistics and uses are not appropriate in a classification problem context.

Ordering and Selecting Variables

Various indices have been used by applied behavioral researchers to order variables in terms of their relative importance to group separation. These ordering indices have also served as bases for selecting good variable subsets; or, to delete poor variables -- "good" and "poor" in the sense of, again, group separation (see Footnote 1). Ordering indices and subset selection methods will be briefly reviewed. A particular index and a particular method will be proposed and advocated for use in group separation problems.

Ordering Indices

One index of variable ordering or importance was already mentioned: the entry order yielded by a stepwise analysis using the F-to-enter value of [7]. The big problem with this index is that the meaning of "importance" is not clear. An F-to-enter considers only the q variables already in the analysis plus one candidate for entry; the remaining p-q-1 variables are ignored. The variable to be entered, i.e., the next most "important" one, is dependent on those already in the analysis; if a different subset of size q was already in, the (q+1)st most "important" variable may very well be different. There are more defensible indices of variable importance.

An index of variable importance that has had some support is the correlation between X_i and the first linear discriminant function (LDF), r_{1i}. A rationale here is that if r_{1i} is "high," then X_i would yield separation comparable to that yielded by the first LDF, the optimal linear

combination. Notwithstanding this rationale, meaning of "importance" still may be a bit vague for some researchers. Also, if more than one LDF is to be retained, an ad hoc index is called for; one that has been proposed is $\sum_{j=1}^{r} = r_{ji}^2$ (see Cooley & Lohnes, 1971, p. 253). [It should be recognized that these LDF-variable correlations are appropriate for another purpose in discrimination problems: to identify or describe the constructs represented by the LDFs that underlie the group separation, i.e., the main or interaction effects.]

A third index is one that has been used with undue frequency: the standardized weight, v_{1i}^*, of the first LDF. [Note: $v_{1i}^* = v_{1i}\sqrt{e_{ii}}$, where e_{ii} is the ith diagonal element of E, the within-groups SSCP matrix; i.e., e_{ii} is the within-groups sum of squares for variable i.] A rationale for this index is based on the fact that scores on the first LDF yield maximum (with respect to any other linear combination) group separation. A big problem with this index is that if the intercorrelation of two variables is "high," the magnitude of the two corresponding weights may be nearly the same, and this common value may be lower than the weight of a "poorer" variable. Furthermore, because of variable collinearity, weights may be quite imprecise.

There is a transform of v_{ji}^*, however, that has some potential as a meaningful index of variable importance. By considering (in, say, a k-group problem) the k centroids in the space of the LDFs, separation attributed to each retained LDF may be specified. For example, if k = 4, the first LDF may distinguish Group 1 from Groups 2, 3, and 4, while the second LDF may distinguish Groups 1 and 2 from Groups 3 and 4. As such, then, it may be informative to order the variables with respect to separation attributed to each LDF. The suggested index of variable importance for variable i and LDF j is

$$u_{ji} = (v^*_{ji})^2 / r^{ii} , \qquad [11]$$

where r^{ii} is the ith diagonal element of the inverse of the witnin-groups correlation matrix. The division by r^{ii}, in effect, "takes into consideration" the problem of collinearity between variable i and the other p - 1 variables. The degree of collinearity for variable i is reflected in the multiple correlation, $R_{(i)}$, between variable i on the one hand and the remaining p - 1 variables on the other hand. And it is the case that

$$r^{ii} = [1 - R^2_{(i)}]^{-1}.$$

The v^* values in [11] are obtainable from SPSS's DISCRIMINANT program under the heading, STANDARDIZED CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS. The r^{ii} values may be obtained a couple different ways. One way is to obtain the within-groups/cells correlation matrix from SAS's GLM program (printed under the heading, PARTIAL CORRELATION COEFFICIENTS FROM THE ERROR SS&CP MATRIX), and invert this matrix using SAS's PROC MATRIX procedure. A second way is to get E from SPSS's MANOVA program, form a diagonal matrix, D, of the positive square roots of the main diagonal elements of E, find the product, $D^{-\frac{1}{2}} E D^{-\frac{1}{2}}$, using PROC MATRIX, and finally find the inverse of this product matrix, again via PROC MATRIX.

There is a fourth index of variable importance that considers all p variables simultaneously, is appropriate as a single index across all dimensions of separation, and for which the meaning of "importance" is fairly clear. This is the F-to-remove value in [8], values of which are obtained with <u>all</u> variables in the analysis. A large F_i, value implies that deleting X_i, will decrease group separation substantially, i.e., will affect grouping variable effects greatly. In fact, for a one-way layout with two groups, it may be shown that F_i, values are proportional to estimates of the decrease in the

inter-group distances when X_i, is deleted, and take response variable collinearity into consideration in a direct way (See Huberty & Smith, in press).

For a one-way layout with two or more groups, F-to-remove, i.e., F_i, values are readily obtained using either SPSS's DISCRIMINANT or BMD's P7M programs (when all of the variables are entered). F-to-remove values are also obtainable for the separate as well as joint factor effects in a higher-way layout by using the BMDP7M program with contrasts--it is necessary to include the statement CONTRAST in the PLOT paragraph.

Selection Methods

One variable selection method was mentioned above: using a cut-off F-to-enter value in a stepwise analysis. Contrary to popular belief, this method does not always yield the best subset of a given size (best in the sense of yielding the smallest Wilks' lambda). Two other methods that have been employed by some behavioral science researchers involve variable orderings yielded via standardized LDF weights and LDF-variable correlations. Variable subsets determined according to either of these ordering indices--arbitrary cut-off values of v_i^* and r_i are employed--are again not optimal.

In the judgment of the current writer, the best way to determine a good subset of variables in a situation involving a one-way MANOVA layout is to use a FORTRAN program described by McCabe (1975). The McCabe program yields the 10 (at most) best subsets of a given size. The subsets are best in the sense of yielding the 10 smallest Wilks' lambdas. There are two reasons why I like the McCabe program: (1) It yields the subset of a given size with the smallest Λ value; and, therefore, the best subset is found; and (2) A number of other good subsets is available to the researcher.

The second reason is important since if for theoretical reasons some variables are better than others, they may be retained by selecting a second or third best subset, but still a good subset. Very often the second or third best subset, of a given size, has a Λ value that isn't much larger than that for the best subset. Or, a researcher may want to retain a subset that includes some or all of the better variables, as determined by F-to-remove values or the u values in [11].

Following is an example of a McCabe output; the 10 best subsets of size 3 in an analysis of a 15-variable problem are:

Λ	Variable Numbers
.7118	5 9 11
.7199	7 12 14
.7229	9 11 14
.7258	5 11 14
.7380	4 5 14
.7406	7 9 11
.7417	5 8 11
.7433	5 7 11
.7435	5 7 14
.7459	5 9 14

[It is interesting to note that the "best" subset determined by a stepwise analysis is that containing variables 5, 11, and 14, actually the fourth best subset of size three. Incidentally, the three top variables, based on F-to-remove values, are 9, 11, and 14.]

In a situation involving a higher-way MANOVA layout, the way to determine a good subset of variables is to use a program by McHenry (1978). Like the McCabe program, the McHenry program is written in FORTRAN, and bases subset selection on Wilks' lambda. Unlike the McCabe program which uses raw data as input, the McHenry program uses matrix input. Matrices inputted are the E and H SSCP matrices. Use of the McHenry program with a

higher-way layout, then, is a three-step procedure: (i) run SAS's GLM program or SPSS's MANOVA program on the original data, making sure to get the E and H matrices as printed output; (ii) punch the E and H matrices (MANOVA will punch E); and (iii) run the McHenry program using E and H as input. In a two-way layout, the McHenry program is used three times, once for each E-H matrix pair. Although this program does not guarantee a best solution, it "seems to find the best solution exceptionally well in practice" (McHenry, 1978, p. 293). The algorithm is a compromise between examining all possible subsets (as does McCabe's) and other procedures which do not guarantee an optimum solution (such as the stepwise). It is an impressively efficient program. Unlike the McCabe program, the McHenry program outputs only the single "best" subset of each size.

Classification Analyses

If a researcher is studying a prediction problem and, therefore, wants to use a classification analysis, the ordering indices and selection methods discussed in this section are not appropriate. For variable ordering, what such a researcher wants is an ordering based on contribution to classification accuracy. For variable selection, what such a researcher wants is a subset that yields the highest proportion of correct classifications. [It should be noted that Wilks' lambda cannot increase as the number of variables is increased; classification accuracy can, however, decrease as the number of variables increases (see Hecker & Wegener, 1978; Jain & Waller, 1978).] Unfortunately, no computer programs are known (by the current writer, at least) to be available that provide either a "good" variable ordering or "best" variable subsets based on classification accuracy. See, however, Henschke and Chen (1974) and Habbema and Hermans (1977).

Classification Hit Rates

Rule Choice

There are really only two concerns when it comes to choosing which form of a normal-based classification rule to use. One concern is which values of estimated prior probabilities of group membership, q_g, to use. Most researchers simply use $q_g = n_g/N$ where n_g is the size of group g, and $N = \sum_{g=1}^{k} n_g$. This is all well and good assuming that the ratio of the sample sizes, $n_1 : n_2 : \ldots : n_k$ is similar to the ratio of the population sizes. Regardless of the sample sizes, the q_g value used in the LCF [5] or in the QCF [6] ought to estimate the proportion of units in the gth population. For example, suppose a researcher is attempting to classify handicapped or exceptional children who are in need of special services as either "learning disabled" or "emotionally disturbed" or "mentally retarded"; even though, say, 25 of each type of children are to be used, q_g values of 0.33 would be inappropriate since it was recently reported that the ratio of population sizes is roughly estimated to be 35:20:45. Although n_g values are not equal, q_g values may be equal. [See Tatsuoka (1971, pp. 225-226) for more on selection of priors.]

The second concern pertains to the choice between a linear rule or a quadratic rule; i.e., between the use of [5] and [6]. Many researchers will, in essence, leave the decision up to a computer program, or the issue is ignored completely. When the issue is not ignored, the typical approach to the decision-making process is to have a computer program calculate a statistic to test the equality of the population covariance matrices--SAS's DISCRIM calculates a Bartlett chi-squared value, while SPSS's DISCRIMINANT calculates a Box F value. There is some evidence (Olson, 1974) that a test

of the equality of the covariance matrices may give quite misleading results. Olson found that for various departures from multivariate normality, Box's test was overly sensitive to covariance heterogeneity. Moreover, Olson (p. 906) points out that simply using a very small value of α does not make Box's test a more useful diagnostic.

Even so, a researcher may be convinced (by the gods?) that the covariance homogeneity condition is untenable. When might this be? There are two situations. One is when the group sizes are quite different. The second is when some of the predictor variables are dichotomous. With either or both of these situations the researcher can use a quadratic rule. Of classification programs in the three packages, BMD, SAS, and SPSS, the only program (at the time of this writing) that yields (correct) quadratic classification results is SAS's DISCRIM.

More will be said on the linear versus quadratic rule concern at the end of the next subsection.

Internal and External Analyses

The classification statistics in [5] and [6] are about as optimal as multivariate normal theory allows; i.e., the expected proportion of correct classification is nearly maximized for the given data. That is, for [5], any other set of linear combinations of the variables will yield a lower proportion of correct classification for the data on hand. When using the observed mean vector, $\bar{X}_g$, and the observed covariance matrix, S, in [5], the proportion of the N units that is correctly (re)classified will be a biased estimator for the true hit rate. In fact, this hit rate is positively biased; i.e., it overestimates the true hit rate (see Michaelis, 1975, p. 230; Frank, Massy, & Morrison, 1965).[2] Such an analysis, which classifies units whose data are used to determine the classification

statistics, is called an internal classification analysis. The amount of bias in an internal analysis is, as might be expected, a function of the number of variables, the number of units, and the degree of group overlap.

There are two general approaches that can be considered to get a better "handle" on the true hit rate. One approach is to use, as in regression analysis, a "shrinkage" formula. For the two-group case such a formula has been derived by McLachlan (1975). Since this formula is restricted to the two-group case, the present discussion will focus on the second approach, an external classification analysis. An external analysis is one in which data on the units to be classified are not used in constructing the classification statistics. Using the statistical computer packages, there are two ways to accomplish an external analysis. One way is to use a hold-out sample via SPSS's DISCRIMINANT program--this is accomplished with a SELECT statement in the DISCRIMINANT paragraph. The user can specify a certain percent of each group, say 60, to be used to construct the rule, and the remaining 40% of the units will be classified using that rule. For large samples, a hold-out analysis should yield a fairly good hit rate estimate.

A second way to accomplish an external analysis is to use the "leave-one-out" procedure proposed by Lachenbruch (1967). With this procedure, one unit is deleted and the classification statistic is formulated on the remaining N - 1 units; then the deleted unit is classified. These steps are carried out N times to determine a hit rate estimate based on classifications of the deleted units. The leave-one-out estimate may be found by specifying JACK (short for jackknife) in the DISC paragraph of BMD's P7M program. Because of an efficient computational algorithm available, such an analysis is not too expensive.

Since most researchers work under limited budgets and, thus, are not able to secure terribly large samples, the current writer favors the leave-one-out method over the hold-out method. [A minimal sample size is here interpreted as the smallest group size being three times the number of variables--roughly, this implies the minimal number of units needed is $N > 3pk$.]

There is, potentially, a practical problem when it comes to obtaining an external analysis using a quadratic rule. None of the packages will provide an external quadratic analysis. I say potentially because there is some evidence (Michaelis, 1973; Huberty & Curry, 1978) that a linear rule will "hold up" better across repeated sampling. The stability of results yielded by a linear rule is greater than results yielded by a quadratic rule when "small" samples are used and when the normality condition is not met. With these findings in mind and with a "pure" prediction problem on hand, one might conclude that in most situations a sound classification analysis is a linear external one.

Assessment of Observed Hit Rates

Results of a classification analysis may be summarized in a "hit table" as follows (for $k = 3$):

		Predicted Group 1	2	3	
Actual Group	1	n_{11}	n_{12}	n_{13}	$n_{1.} = n_1$
	2	n_{21}	n_{22}	n_{23}	$n_{2.} = n_2$
	3	n_{31}	n_{32}	n_{33}	$n_{3.} = n_3$
		$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..} = N$

A question naturally arises: Are the observed results better than what can be expected by chance? More specifically: Are the k separate-group hit rates, n_{gg}/n_g, and the total-group hit rate, $(\sum_{g=1}^{k} n_{gg})/N$, better than what can be expected by chance? To answer this question, two concerns must be dealt with: (1) the analysis used to yield the observed hit table; and (2) the meaning of "chance." The appropriate analysis to consider, for reasons discussed in the prior subsection, is an external analysis. For total-group hit rates, two interpretations of chance are applicable in classification analyses; these are presented next.

Proportional chance criterion. Consider a situation with two populations of approximately the same size, with samples of a common size selected from each population. It is reasonable to expect, then, that by using the flip of a fair coin to decide group membership, 50% of the units could be classified correctly. Similarly, with three populations and samples of a common size, it is expected that one-third of the units could be correctly classified by chance. In general, with k populations and samples of a common size, it may be expected that 1/k of the units could be correctly classified by chance.

The above conclusion may be justified in the following manner. If predicted group membership is independent of actual group membership, then

the expected frequency in cell (g, g) of a hit table is

$$e_g = \frac{n_g \, n_{\cdot g}}{N} \quad .$$

If $n_1 = n_2 = \ldots = n_k = n$, then the overall chance frequency of hits would be

$$e = \sum_{g=1}^{k} e_g = \sum_{g=1}^{k} \frac{n n_{\cdot g}}{N} = n.$$

Thus, the chance hit rate would be $n/N = 1/k$.

Now consider a situation where the k populations are of varying sizes, and the sample sizes reflect this variation. It is then reasonable to assume that the proportions, $q_g = n_g/N$, yield good estimates of the prior probabilities of group membership (also called base rates). If so, then the chance frequency of hits for group g is

$$e_g = \frac{n_g \, n_{\cdot g}}{N} = q_g \, n_{\cdot g} \quad . \qquad [12]$$

The total-group chance frequency of hits is

$$e = \sum_{g=1}^{k} q_g \, n_{\cdot g} \quad ,$$

and the total-group chance hit rate is

$$\frac{e}{N} = \frac{1}{N} \sum_{g=1}^{k} q_g \, n_{\cdot g} = \sum_{g=1}^{k} \frac{n_g \, n_{\cdot g}}{N} \quad .$$

This is termed the proportional chance criterion.

To answer the question posed at the outset of this subsection for the total-group hit rate, then, the value of e is compared with the observed total frequency of hits:

$$o = \sum_{g=1}^{k} n_{gg} \quad .$$

From a statistical significance point of view, the question may be expressed as: "Is the number of units correctly classified significantly greater than the number correctly classified by chance?" When the overall number of hits, o, is less than or equal to the chance number, e, it is not necessary to test for significance. Hence the significance test involves a directional alternative hypothesis.

Now, o may be considered a binomial statistic that can take on values from zero to N. In classification problems, N is generally large enough so that a normal probability distribution may be used to approximate the distribution of o. Thus, the standardized normal statistic,

$$\frac{(o - e)\sqrt{N}}{\sqrt{e(N - e)}}, \qquad [13]$$

may be used as a test statistic. The lower bound of a "one-sided" interval estimate of the true overall frequency of hits may be determined using

$$o - z_{1-\alpha}\sqrt{e\,(N - e)/N} \quad .$$

[It should be noted that it is inappropriate to answer the better-than-chance question by applying the Pearson chi-squared statistic to a hit table. This is because only the diagonal entries are of interest as far as the significance question is concerned. If the observed hits yield significance, so too would the Pearson statistic. However, a significant Pearson statistic does not imply a better-than-chance result. That is, a significant Pearson statistic is necessary but not sufficient to conclude that a rule yields better than chance classification.]

Maximum chance criterion. Consider now a second meaning of "chance classification." Suppose the following hit table resulted from a classification analysis:

		Predicted Group 1	2	3	
Actual Group	1	115	20	15	150
	2	2	20	3	25
	3	1	4	20	25
		118	44	38	200

Using the proportional chance criterion we would expect to correctly classify, by chance, $e = 98.75$ units out of the 200. The observed hit total of $o = 115 + 20 + 20 = 155$ (a hit rate of 0.775) yields a standardized normal statistic value (see [13]) of 7.96 ($P << .001$). With such results one might be tempted to conclude that the employed classification rule was acceptably accurate.

Consider, however, a researcher who states, "One could get a hit rate of 0.75 by simply assigning all of the 200 units to Group 1." Such an interpretation of "chance" leads to the maximum chance criterion, $\max_g (n_g/N) = .75$, which implies that 75% of the units could be correctly classified by chance. Using this as the expected hit rate and the results from the above table, the standardized normal statistic value (see [13]) is 0.816 ($P > .20$). As such, one might question whether the rule used to get the results in the above table yielded classification accuracy better than "chance."

Separate-group hit rates. There are research situations in which hit rates for specific groups are of particular interest. Relative group sizes in the following examples are approximately the same as those used in the

reported studies. The results, however, were modified somewhat for purposes of discussion herein. Example 1 (based on Elkins & Sultmann, 1981): Here, the problem is classifying children, on the basis of psycholinguistic ability data profiles, as either learning disabled (LD) or "normal" (N). A resulting hit table might be:

		Predicted Group: LD	Predicted Group: N	
Actual Group	LD	30	50	80
	N	40	280	320
		70	330	400

Example 2 (based on Gully & Hosch, 1979): Here, the problem is classifying children, on the basis of adaptive behavior measures, as nonretarded (NR), educable mentally retarded (EMR), or trainable mentally retarded (TMR). Suppose the following hit table resulted:

		Predicted Group: TMR	Predicted Group: EMR	Predicted Group: NR	
Actual Group	TMR	65	20	0	85
	EMR	45	25	45	115
	NR	15	75	320	400
		125	120	365	600

Example 3 (based on LaRocco, Pugh, & Gunderson, 1977): Here, the problem is predicting, on the basis of demographic and performance measures, whether Navy enlisted men will reenlist (RR), not reenlist (RN), or separate from service (NN). Consider the following hit table:

	Predicted Group RR	NN	RN	
RR	35	35	5	75
NN	5	70	25	100
RN	50	30	420	500
	90	135	450	675

Using the maximum chance criterion, it cannot be concluded in any of the three examples that classification accuracy is better than "chance." A more interesting assessment, however, may pertain to separate-group classification accuracy. In terms of statistical significance, the standardized normal test statistic for group g is

$$\frac{(n_{gg} - e_g)\sqrt{n_g}}{\sqrt{e_g\,(n_g - e_g)}} \qquad [14]$$

A question of particular interest in Example 1 may pertain to how accurately LD children can be identified. Here, $e_1 = 80(70)/400 = 14$ (see [12]), and the value of [14] is $z = 9.56$ ($P << .001$). In Example 2, TMR children can be accurately identified ($z = 12.63$, $P << .001$), while EMR children cannot ($z = .47$, $P > .30$). In Example 3, both RR and NN types can be accurately predicted--$z = 8.49$, $P << .001$ and $z = 12.50$, $P << .001$, respectively.

In a situation where the proportional chance criterion is appropriate for assessing overall classification accuracy, statistic [14] may also be employed in examining the goodness of separate-group hit rates.

Improvement over chance. A value of statistics [13] or [14] yields one assessment of the improvement over chance of an observed hit rate. There is, however, an alternative assessment procedure. A researcher may be interested in determining how much better he/she can predict group membership using a

classification rule over that using some "chance" assignment. An improvement-over-chance or a reduction-in-error index useful for such a purpose is

$$I = \frac{H_o - H_e}{1 - H_e} \quad ,$$

where H_o is the observed hit rate, and H_e is the hit rate expected by "chance." The total-group chance hit rate may be obtained using either of the two criteria discussed above. The index, I, is a proportional-reduction-in-error statistic in that 100I% fewer classification errors result using a classification rule than would be expected by chance classification. For example, with the hypothetical three-group results (with N = 200) given earlier in this subsection, I is equal to 0.56, using the proportional chance criterion; thus, it could be concluded that 56% fewer errors would be made than would be expected by chance. Using the maximum chance criterion, I = .10.

As an example of using I for a separate-group hit rate, consider Example 2 above. For the TMR group, $H_o = 65/85 \doteq .76$, $H_e \doteq .21$, and $I \doteq .70$; thus, about 70% fewer errors would be made in identifying TMR children using the available data than would be expected by chance.

Klecka (1980, p. 51) suggested this index in a classification context using the proportional chance criterion. In a different context, Guttman (1941) proposed the index (using, essentially, the maximum chance criterion) some 40 years ago. The I index may also be recognized as a version of Cohen's (1960) kappa, which was recently discussed by Brennan and Prediger (1981).

As an alternative to the inferential assessment made using [13] or [14], one can test for statistical significance of I (when H_e is based on the proportional chance criterion) using the estimated standard error given by

Fleiss, Cohen, and Everitt (1969). To complicate matters somewhat, it may be reasonable to incorporate differential costs of classification errors and employ a "weighted I" index (see, again, Fleiss et al., 1969).

Threshold Probabilities

It was mentioned in an early section of this paper that a classification statistic may be expressed as an estimate of a posterior probability of group membership, $\hat{P}(G_g|\underline{X})$ (see expression [2]). For a given unit, k values of $\hat{P}(G_g|\underline{X})$ are determined, and the unit is assigned to that group which yields the largest value. [All three package programs yield estimated posterior probabilities, with some variations; SPSS's DISCRIMINANT outputs only the two highest values; with the JACK option, BMD's P7M printed values will be a little different from $\hat{P}(G_g|X)$.] In a true prediction problem, where a goal is to determine a classification rule, it is assumed that the units employed originate from k well-defined meaningful populations, and that any subsequent unit whose classification is to be determined also originates from one of the same populations. It may be decided, however, that some "new" units should not be assigned to any of the original populations because $\hat{P}(G_g|\underline{X})$ values do not make such membership clear. The basis for such a decision may be a minimum, or threshold, $\hat{P}(G_g|\underline{X})$ value. The following estimated probabilities were found with real data: $\hat{P}(G_1|\underline{X}) = .003$, $\hat{P}(G_2|\underline{X}) = .493$, $\hat{P}(G_3|\underline{X}) = .505$. It is not clear whether the unit should be assigned to Group 2 or to Group 3. If a $\hat{P}(G_g|\underline{X})$ value is not "significantly" greater than, say, 0.50, then it may be decided that the unit should be put in a category of "cannot be classified." SAS's DISCRIM program has an option that allows the user to specify a threshold value. The choice of a threshold value is an arbitrary one, and would be dependent on the overall sample size, the number of groups, and the number of predictors.

Classification and Discrimination

As mentioned at the outset of this paper, a basic purpose of predictive discriminant analysis techniques involves classifying or assigning units to well-defined groups or populations. Classification techniques can also be useful in one-way layout group-separation, or discrimination, problems. One use of classification techniques in discrimination problems is very common--very common, perhaps, because output from two popular computer programs, SPSS's DISCRIMINANT and BMD's P7M, include PDA results as well as DDA results. This use involves using classification results to supplement and corroborate discrimination results. A high degree of classification accuracy supports a finding of little overlap among the groups in a sort of "descriptive" sense. See, for example, Tankard and Harris, (1980), Terenzini, Hartmark, Lorang, and Shirley (1980), and Elkins and Sultmann (1981).

There are three potential uses of classification results/techniques in group separation problems, none of which have been common. One is the examination of estimated posterior probabilities of group membership, $\hat{P}(G_g|\underline{X})$, to detect units that are "fence riders," and of estimated probabilities, $\hat{P}(\underline{X}|G_g)$, which are proportional to estimates of $f(\underline{X}|G_g)$ in [2], to detect units that are "outliers." [For a unit assigned to group g, the single $\hat{P}(\underline{X}|G_g)$ value is printed by SPSS's DISCRIMINANT.] The study of clusters of fence-riders and outliers may very well reveal interesting effects, in a "descriptive" sense, of the grouping variable. See, again, Tankard and Harris (1980), and Terenzini et al. (1980).

Another use of classification techniques is the study of group separation in a one-way layout when it is clear that the covariance homogeneity condition is not met. In such a situation, degree of group overlap can be

assessed, again in a "descriptive" sense, by examining the results of a quadratic classification analysis.

Finally, classification ideas can be employed in designing a two-group separation study. Lachenbruch (1968) developed tables of sample sizes needed to yield a classification hit rate within a given tolerance of an optimum hit rate. This approach is described in detail by Huberty and Holmes (in press) for a univariate problem--that paper also illustrates the usefulness of examining individual unit proabability estimates to identify fence-riders and outliers.

Summary

There are many, many issues in the interpretation of discriminant analyses, only a few of which are discussed in this paper. The issues presented are restricted to those pertaining to typical uses of discriminant analysis by researchers in the behavioral sciences. Because many behavioral researchers rely on the use of the popular computer program packages, most of the issues discussed deal with information obtainable from the package discriminant analysis programs. A summary of the issues reviewed in this paper is presented in the form an an annotated list.

1. Distinguishing between a linear discriminant function (LDF) and a linear classification function (LCF). Although both are linear composites of variables, LDFs and LCFs are arrived at differently, are different in number, are interpreted differently, and are used for different purposes. LDF scores may be used in classification problems; if so, classification decisions are typically not based on these scores, but rather on LCF scores with LDF scores as input.

2. Misusing stepwise discriminant analysis programs. Stepwise programs are often used in a discrimination problem to determine the relative

importance of variables (i.e., variable ordering) and what variables should be retained for interpretation (i.e., variable selection), and to order and select variables in a classification problem. The position taken by the current writer is that stepwise programs should not be used for any of these purposes.

3. Ordering variables and selecting variable subsets. To order variables in terms of specific group separation, an index which is a transform of a standardized LDF weight is suggested. A defensible index of relative variable importance, which yields a single ordering for a given effect, is the F-to-remove value. An excellent computer program to use to determine a good variable subset for a one-way layout is one by George McCabe (Purdue), and for a two-way layout, a program by Claude McHenry (Tennessee Tech).

4. Choosing a classification rule. The prior probabilities of group membership should reflect the relative sizes of the k populations involved. Except for special situations, a linear rule should be used.

5. Estimating true classification hit rates. An external analysis should be used: the hold-out sample approach if sample size is large enough; otherwise, and more often, the leave-one-out approach.

6. Assessing classification accuracy. If group sizes are near equal, the proportional chance criterion is appropriate. If sample sizes are quite diverse, the maximum chance criterion is to be preferred. [The Pearson chi-squared statistic is inappropriate in any case.] Separate-group hit rates should not be overlooked, particularly when group sizes are quite different. An improvement-over-chance index is suggested.

7. Examining and using classification results. Estimated probabilities for individual units can and should be considered both in classification problems (using threshold probabilities in assignment decisions) and in

group separation problems (using probability estimates to identify fence-riders and outliers). Classification results may also be utilized in a descriptive way to corroborate group separation findings.

References

Afifi, A.A., & Azen, S.P. Statistical analysis. New York: Academic Press, 1979.

Anderson, T.W. Introduction to multivariate statistical analysis. New York: Wiley, 1958.

Brennan, R.L., & Prediger, D.J. Coefficient kappa: Some uses, misuses, and alternatives. Educational and Psychological Measurement, 1981, 41, 687-699.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cooley, W.W., & Lohnes, P.R. Multivariate data analysis. New York: Wiley, 1971.

Elkins, J., & Sultmann, W.F. ITPA and learning disability: A discriminant analysis. Journal of Learning Disabilities, 1981, 14, 88-92.

Fleiss, J.L., Cohen, J., & Everitt, B.S. Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 1969, 72, 323-327.

Frank, R.E., Massy, W.F., & Morrison, D.G. Bias in multiple discriminant analysis. Journal of Marketing Research, 1965, 2, 250-258.

Gondek, P.C. What you see may not be what you think you get: Discriminant analysis in statistical packages. Educational and Psychological Measurement, 1981, 41, 267-282.

Gully, K.J., & Hosch, H.M. Adaptive Behavior Scale: Development as a diagnostic tool via discriminant analysis. American Journal of Mental Deficiency, 1979, 83, 518-523.

Guttman, L. An outline of the statistical theory of prediction. In P. Horst (Ed.), The prediction of personal adjustment. New York: Social Science Research Council, 1941.

Habbema, J.D.F., & Hermans, J. Selection of variables in discriminant analysis by F-statistic and error rate. Technometrics, 1977, 19, 487-493.

Hecker, R., & Wegener, H. The validation of classification rates in stepwise discriminant analyses. Biometrical Journal, 1978, 20, 713-727.

Henschke, C.I., & Chen, M.M. Variable selection technique for classification problems. Educational and Psychological Measurement, 1974, 34, 11-18.

Huberty, C.J, & Curry, A.R. Linear versus quadratic multivariate classification. Multivariate Behavioral Research, 1978, 13, 237-245.

Huberty, C.J, & Holmes, S.E. Two-group comparisons and univariate classification. Educational and Psychological Measurement, in press.

Huberty, C.J, & Smith, J.D. The study of MANOVA effects. Multivariate Behavior Research, in press.

Jain, A.K., & Waller, W.G. On the optimal number of features in the classification of multivariate Gaussian data. Pattern Recognition, 1978, 10, 365-374.

Jennrich, R.I. Stepwise discriminant analysis. In K. Enslein et al. (Eds.), Statistical methods for digital computers. New York: Wiley, 1977.

Klecka, W.R. Discriminant analysis. Beverly Hills, Calif.: Sage, 1980.

Lachenbruch, P.A. An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics, 1967, 23, 639-645.

Lachenbruch, P.A. On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficient. Biometrics, 1968, 24, 823-834.

La Rocco, J.M., Pugh, W.M., & Gunderson, E.K. Identifying determinants of retention decisions. Personnel Psychology, 1977, 30, 199-215.

McCabe, G.P. Computations for variable selection in discriminant analysis. Technometrics, 1975, 17, 103-109.

McHenry, C.E. Computation of a best subset in multivariate analysis. Applied Statistics, 1978, 27, 291-296.

McLachlan, G.J. Confidence intervals for the conditional probability of misallocation in discriminant analysis. Biometrics, 1975, 31, 161-167.

Michaelis, J. Simulation experiments with multiple group linear and quadratic discriminant analysis. In T. Cacoullos (Ed.), Discriminant analysis and applications. New York: Academic Press, 1973.

Morrison, D.F. Multivariate statistical methods. New York: McGraw-Hill, 1976.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., & Bent, D.H. Statistical package for the social sciences. New York: McGraw-Hill, 1975.

Olson, C.L. Comparative robustness of six tests in multivariate analysis of variance. Journal of the American Statistical Association, 1974, 69, 894-908.

Tankard, J.W., & Harris, M.C. A discriminant analysis of television viewers and nonviewers. Journal of Broadcasting, 1980, 24, 399-409.

Tatsuoka, M.M. Multivariate analysis. New York: Wiley, 1971.

Terenzini, P.T., Hartmark, L., Lorang, W.G., & Shirley, R.C. A conceptual and methodological approach to the identification of peer institutions. Research in Higher Education, 1980, 12, 347-364.

Footnotes

[1]Henceforth in this paper, the terms "separation" and "group separation" will be used in the context of a discrimination problem, can be read with either a one- or higher-way layout in mind, and are meant to cover interaction effects as well as main effects.

[2]This result is similar to that in multiple regression analysis where the observed correlation coefficient, R, is a positively biased estimator.